Large Deviation Delay Analysis of Queue-Aware Multi-user MIMO Systems with Multi-timescale Mobile-Driven Feedback

Junting Chen, *Student Member, IEEE* and Vincent K. N. Lau, *Fellow, IEEE*
Dept. of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
Email: {eejtchen, eeknlau}@ust.hk

Abstract

Multi-user multi-input-multi-output (MU-MIMO) systems transmit data to multiple users simultaneously using the spatial degrees of freedom with user feedback channel state information (CSI). Most of the existing literatures on the reduced feedback user scheduling focus on the throughput performance and the user queueing delay is usually ignored. As the delay is very important for real-time applications, a low feedback queue-aware user scheduling algorithm is desired for the MU-MIMO system. This paper proposed a two-stage queue-aware user scheduling algorithm, which consists of a queue-aware mobile-driven feedback filtering stage and a SINR-based user scheduling stage, where the feedback filtering policy is obtained from the solution of an optimization problem. We evaluate the queueing performance of the proposed scheduling algorithm by using the sample path large deviation analysis. We show that the large deviation decay rate for the proposed algorithm is much larger than that of the CSI-only user scheduling algorithm. The numerical results also demonstrate that the proposed algorithm performs much better than the CSI-only algorithm requiring only a small amount of feedback.

Index Terms

MU-MIMO, Limited Feedback, Queue-aware, Large Deviation, Random Beamforming

Large Deviation Delay Analysis of Queue-Aware Multi-user MIMO Systems with Multi-timescale Mobile-Driven Feedback

I. Introduction

MIMO is an important core technology for next generation wireless systems. In particular, in multi-user MIMO (MU-MIMO) systems, a base station (BS) (with M transmit antennas) communicates with multiple mobile users simultaneously using the spatial degrees of freedom at the expense of knowledge of channel states at the transmitter (CSIT). It is shown in [1], [2] that using simple zero-forcing precoder and near orthogonal user selection, a sum rate of $M \log\log K$ can be achieved with full CSIT knowledge. Yet, full CSIT knowledge is difficult to achieve in practice and there are a lot of works focusing on reducing the feedback overhead in MIMO systems [3]–[8]. For instance, in [3], [4], the authors have focused on the codebook design and performance analysis under limited-rate feedback schemes. In [5]–[7], on the other hand, a threshold based feedback control is adopted where users attempt to feedback only when its channel quality exceeds a threshold. It was further shown that a sum rate capacity $\mathcal{O}(M \log\log K)$ can be achieved when only $\mathcal{O}(M \log\log\log K)$ users feeding back to the BS [5].

While there are a lot of works that consider reduced feedback design for MU-MIMO, all these existing works focused on the throughput performance. They have assumed infinite backlog at the base station and therefore, ignored the bursty arrival of the data source as well as the associated delay performance, which is very important for real-time applications. For instance, the CSI information indicates *good opportunity to transmit* whereas the *Queue State Information* (QSI) indicates the *urgency* of the data flow. A delay-aware MU-MIMO system should incorporate both the CSI and QSI in the user scheduling. However, it is far from trivial to integrate these information in determining the user priority. There are some works considering QSI in the user scheduling of MU-MIMO systems. In [9], the author considered a queue-aware power control and dynamic clustering in downlink MIMO systems. In [10], the authors considered MU-MIMO user scheduling to maximize queue-weighted sum rate. Due to the exponentially large solution space, heuristic greedy-based algorithm is proposed. However, these works required the BS to have global CSI knowledge of all the users, which is hard to achieve in practice. Furthermore, the delay performance in [10] is obtained by simulation only and not much design insights can be obtained in these works. In general, there are still a number of first order technical challenges associated with designing delay-aware MU-MIMO systems.

- **Challenges in User Scheduling Design**: For real-time applications, it is important to exploit CSI and QSI in the user scheduling. Yet, it is highly non-trivial to design a *priority metric* that strike a balance between transmission opportunity and urgency. A brute-force stochastic optimization approach such as MDP [11], [12] will end up having huge complexity solution (exponential w.r.t. K), which is highly impractical. On the other hand, brute-force application of Lyapunov optimization techniques [13] in MU-MIMO is also not feasible because of the associated exponential complexity of user selection for MU-MIMO.
- **Challenges in Delay Analysis**: Due to the QSI-aware control algorithm, the service rate of the data queues are *state-dependent* and the queue dynamics from these K data flows are coupled together. This makes the queueing delay analysis extremely difficult. There is no closed form results on the steady state distributions of the queue length in such complex queueing systems. In [14], the authors characterized the *stability region* of the MU-MIMO systems under limited CSI feedback. Yet, stability is only a weak form of delay performance.

In this paper, we consider a MU-MIMO downlink system with a M-antenna BS and K multi-antenna mobile users. The BS applies the random beamforming for MU-MIMO to exploit the multi-user diversity. To overcome the complexity challenge of user scheduling, we propose a two timescale delay-aware user scheduling policy for the MU-MIMO system. The proposed policy consists of two stages, namely the *queue-aware user-driven feedback filtering stage* and the *dynamic SINR-based user scheduling stage*. At the first stage (slower timescale), the BS broadcasts a QSI-dependent user feedback candidate list and only mobiles in the list are allowed to feedback the CSI to the BS. At the second stage (faster timescale), the BS selects the strongest users based on the CSI of the users selected in the first stage. Based on the two timescale user scheduling policy, we then analyze the delay performance of the MU-MIMO system. It is in general difficult to analyze the delay for state-dependent coupled queues. To overcome this challenge, we consider the large deviation tail for the maximum queue length among all the users, which reflects the worse case delay performance in the system. Using large deviation theory for random process [15], we derive the asymptotic exponential decay rate for the tail probability of the maximum queue length. Specifically, we quantify the asymptotic decay rate $-\frac{1}{B}\log(\mathrm{P}(\max_k\ Q_k) > B)$ as the buffer size $B \to \infty$. We show that the decay rate of the worst case queue length of the proposed delay-aware scheduling algorithm scales as $\mathcal{O}(\log K)$, which is substantially better than traditional MU-MIMO user scheduling baseline schemes.

The rest of the paper is organized as follows. We present the system model, bursty data source and queueing model and the proposed two timescale delay-aware user scheduling policy in Section II. In Section III, we derive the optimal user-driven feedback filtering strategy using Lyapunov approach. We then analyze the maximum queue length property using sample path fluid approximation and

large deviation theory in Section IV. Numerical results are provided in Section V and we conclude the results in Section VI.

Notations: $f(x) = \mathcal{O}(g(x))$ denotes $\lim_{x\to\infty} \frac{f(x)}{g(x)} < \infty$, $f(x) = o(g(x))$ denotes $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$, and $\mathbb{E}_x[f(x,y)] = \int f(x,y)dF_x(x)$ denotes the expectation over random variable x (treating y as constant).

II. System Model

A. MU-MIMO System Model

We consider a downlink MU-MIMO system with a M-antenna BS and K geometrically dispersed mobile users ($K \gg M$). Each mobile user has N receive antennas. Using MU-MIMO techniques, the BS transmits M data streams to a group of selected users at each time slot. The wireless channel between each user and the BS is modeled as a Rayleigh fading channel. Specifically, the received signal $\mathbf{y}_k \in \mathbb{C}^{N\times 1}$ by the user k is given by

$$\mathbf{y}_k = \sqrt{P}H_k\mathbf{x} + \mathbf{n}_k \qquad \forall k \in \mathcal{A}(t) \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is the normalized transmitted signal with $\mathbb{E}\left[\text{Tr}(\mathbf{x}\mathbf{x}^*)\right] = M$, i.e., the normalized transmit power on each antenna is assumed to be one, $H_k \in \mathbb{C}^{N\times M}$ is the zero mean, unit-variance circularly symmetric complex Gaussian channel matrix from the transmitter to the user k, $\mathbf{n}_k \in \mathbb{C}^{N\times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ is the Gaussian additive noise vector, P is the transmit power at the BS, and $\mathcal{A}(t)$ denotes the set of the scheduled users at time slot t. We have the following assumption on the channel matrices $\{H_k\}$.

Assumption 1 (Assumptions on Channel Matrices): The channels are assumed to be in *quasi-static* block fading, where each channel realization H_k remains constant during each time slot, but identically and independently distributed (i.i.d.) across different time slots. The mobile users are assumed to have perfect knowledge of their local CSI. However, only a selected portion of the users will feedback their CSI to the BS and the feedback information is delivered through a noiseless feedback channel. ■

At the BS, random beamforming is used to support near-orthogonal data streams transmissions to the selected users without knowing the full CSI $\{H_k\}$. The BS chooses M random orthonormal vectors $\{\phi_1, \ldots, \phi_M\}$, where $\phi_m \in \mathbb{C}^{M\times 1}$ are generated according to an isotropic distribution. Let $\mathbf{s}(t) = (s_1(t), \ldots, s_M(t))$ be the vector of the transmit symbols. The transmit signal is given by

$$\mathbf{x}(t) = \sum_{m=1}^{M} \phi_m s_m(t).$$

Therefore, the receive signal at the k-th user is

$$\mathbf{y}_k(t) = \sum_{m=1}^{M} \sqrt{P}H_k\phi_m s_m(t) + \mathbf{n}_k.$$

We assume the receivers know the beamforming vectors $\{\phi_m\}$. The *effective SINR* of the i-th beam on the n-th receive antenna of the k-th user can be calculated as follows,

$$\text{SINR}_{k,n}^{i} = \frac{\left|H_k^{(n)}\phi_i\right|^2}{\sum_{j,j\neq i}\left|H_k^{(n)}\phi_j\right|^2 + 1/P}. \tag{2}$$

where $H_k^{(n)}$ denotes the n-th row of the channel matrix H_k of user k. By selecting the users with the highest SINR on each beam, the transmitter can support near-orthogonal transmission and exploit multi-user diversity without the global CSI $\{H_k\}$ [16].

B. Bursty Data Source and Queue Model

Data arrives in packets randomly for different users. Let $A_k(t)$ denote the number of packets that arrive at the BS for user k during time slot t, and $\mathbf{A}(t) = (A_1(t), \ldots, A_k(t))$. We assume that the arrivals $A_k(t)$ are i.i.d over different time slot t. We have the following assumptions regarding the bursty arrival processes $A_k(t)$.

Assumption 2 (Bursty Source Model): The packet arrival $A_k(t)$ is i.i.d. with respect to (w.r.t.) t and independent w.r.t. k according to a general distribution with mean $\mathbb{E}[A_k(t)] = \lambda_k$ and moment generating function (MGF) $\Lambda_{A,k}(\theta) = \mathbb{E}\left[e^{\theta A_k}\right]$. The packet length is assumed to be constant L bits. ■

The BS maintains queueing backlogs $Q_k(t)$ for each user k. Let $D_k(\mathbf{Q}(t), \mathbf{H}(t))$ represents the amount of departure for user k at time slot t, where $\mathbf{Q}(t) = (Q_1(t), \ldots, Q_K(t))$ and $\mathbf{H}(t) = (H_1(t), \ldots, H_K(t))$. The queueing dynamics for user k is given by

$$Q_k(t+1) = [Q_k(t) - D_k(\mathbf{Q}(t), \mathbf{H}(t))]^+ + A_k(t) \tag{3}$$

where the operator $[\centerdot]^+$ represents $[w]^+ = \max\{0, w\}$. Using Little's Law [17], the average delay of the k-th user is given by $\overline{T}_k = \overline{Q}_k/\overline{D}_k$, where $\overline{Q}_k$ is the average backlog for the k-th queue and $\overline{D}_k$ is the average departure at each time slot. As a result, there is no loss of generality to study the queue length Q_k for the purpose of understanding the delay. Obviously, the queue length (or the delay) of the MU-MIMO system depends on how we use the channel resources. Hence the goal of the user scheduling controller is to adjust the channel access opportunity for all the users so that their queue lengths (or delay) are minimized while a reasonable system throughput is maintained.

C. Two Timescale User Scheduling with Reduced Feedback for MU-MIMO

A reasonable delay-aware user scheduling algorithm should jointly adapt to both the CSI (to capture good transmission opportunity) and the QSI (to capture the urgency). In particular, we are interested in the control policy that can maximize queue stability region. However, conventional throughput

optimal (in stability sense) user scheduling policies such as max-weighted-queue (MWQ) algorithms [13], [14] require global CSI and QSI knowledge. However, the CSI is available at the mobile user side while the QSI is available at the BS. Furthermore, the MWQ policy requires solving a queue weighted sum rate combinatorial optimization problem, which has exponential searching space. Hence, brute force solution of the MWQ problem requires huge signaling overhead as well as huge complexity. To overcome these challenges, we propose a two timescale user scheduling solution as follows.

- Stage I: *Queue-aware user-driven feedback filtering*. The BS determines and broadcasts the user feedback probability $\{p_1(\mathbf{Q}), \ldots p_K(\mathbf{Q})\}$ based on the user queueing backlogs $\mathbf{Q}(t)$ for every T time slots. Mobile user k attempts to feedback to the BS in the stage II with probability p_k. We denote $\chi_k \in \{0, 1\}$ as the stochastic feedback filtering policy with $\mathrm{P}(\chi_k = 1) = p_k$, and a user k feeds back when $\chi_k(t) = 1$. The motivation of the mobile feedback filtering is to save the feedback cost by reducing the lower priority users from feeding back.
- Stage II: *Dynamic User Scheduling based on SINR feedbacks*. If the feedback filtering policy $\chi_k = 1$, then user k measures the effective SINR vector $\{\mathrm{SINR}_{k,n}^1, \ldots, \mathrm{SINR}_{k,n}^M\}$ on each receive antenna n according to (2) and finds the strongest beam $i^*(k,n) = \arg\max_{1 \le i \le M} \mathrm{SINR}_{k,n}^i$. The mobile then feeds back the selected beam index $i^*(k,n)$ and the associated $\mathrm{SINR}_{k,n}^{i^*(k,n)}$ to the BS. The set of feedback users at time slot t is denoted by $\mathcal{F}(t)$. The BS schedules user $k^*(i)$ to transmit at the i-th beam who has the highest SINR, i.e., $k^*(i) = \arg\max_{k \in \mathcal{F}(t)} \gamma_k^i$, where $\gamma_k^i = \max_{n \in \mathcal{N}(k,i)} \mathrm{SINR}_{k,n}^i$ denotes the highest SINR of user k on the i-th beam. Here $\mathcal{N}(k,i) = \{n : 1 \le n \le N, i^*(k,n) = i\}$ denotes the set of antennas of user k which have fed back the SINR for the i-th beam[1]. As a result, the stage II user scheduling exploits the multi-user diversity among the set of users attempting to feedback $\mathcal{F}(t)$.

Fig. 1 depicts an illustration of the two stages user scheduling policy. The policy tries to balance the transmission opportunity and urgency with a low complexity and low feedback cost strategy. For the user with a long queue, it will be given priority to feedback during the stage I feedback filtering phase. Users who have passed the stage I filtering will compete for channel access based on the stage II SINR based scheduling in which users with better channel conditions will be served. Moreover, the two stages processing can be implemented on different timescale. The user selection in stage II is done at every time slot t, while the user feedback probability $\{p_k(\mathbf{Q})\}$ determined in stage I can be updated once every T time slots. The update period T trades the performance of the two timescale policy with the control signaling overhead. With a large T, there is a smaller signaling overhead associated with broadcasting $\{p_k(\mathbf{Q})\}$ in stage I but the feedback priority may be driven by outdated QSI.

[1]We define $\gamma_k^i = 0$ if $\mathcal{N}(k,i) = \emptyset$.

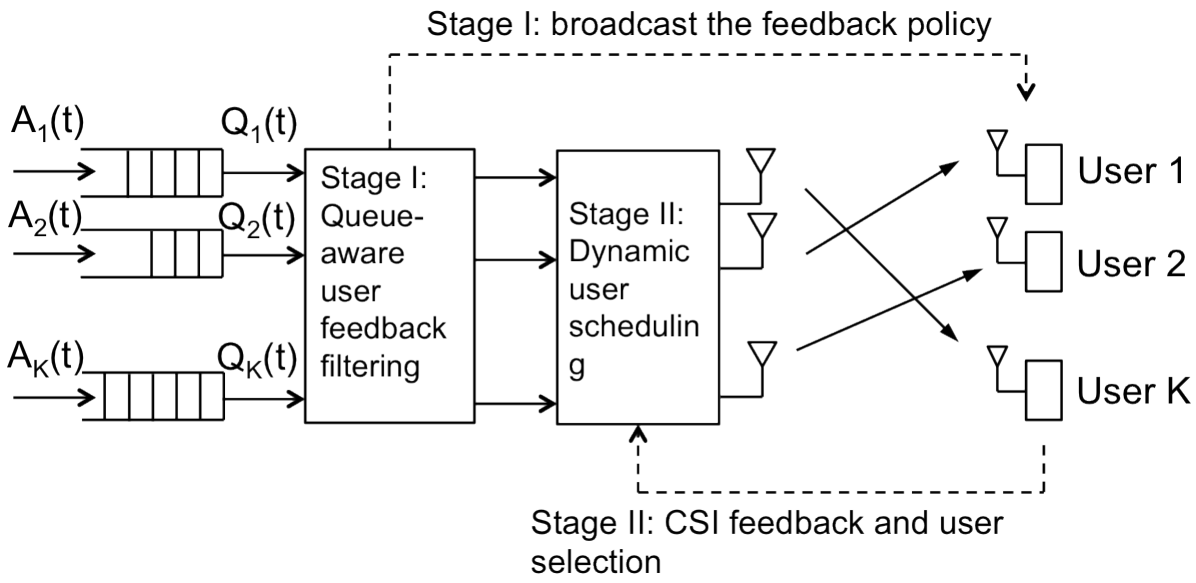

Figure 1. The two stage joint CSI and QSI user scheduling in a multi-user MIMO system. At stage I, the BS determines the user feedback priority based on the QSI. At stage II, a portion of selected users feedback their CSI and the BS schedules users for transmission based on their CSI feedback.

D. Queue-Aware Feedback Filtering (Stage I) Optimization

The feedback filtering control in stage I plays a critical role in the overall delay performance of the MU-MIMO system. In the following, we adopt a Lyapunov optimization technique to derive the stage I feedback filtering policy to achieve the maximum *queue stability region* in the MU-MIMO system.

1) Queue Stability : We first define the queue stability and the stability region formally below.

Definition 1 (Queue Stability): The queueing system is called *stable* if

$$\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}\left[\max_k Q_k(t)\right]<\infty.$$

■

Definition 2 (Stability region and Throughput Optimal Policy): The *stability region* $\mathcal{C}$ is the closure of the set of all the arrival rate vectors $\{\lambda_k\}$ that can be stabilized in a MU-MIMO system, using the two timescale user scheduling (Stage I and Stage II) framework in Section II-C. A *throughput optimal* user scheduling policy is a policy that stabilizes all the arrival rate vectors $\{\lambda_k\}$ within the stability region $\mathcal{C}$. ■

2) The Data Rate and the Amount of Feedback: Let $J_k^i \in \{0,1\}$ be the scheduling indicator of the k-th user on the i-th beam. Note that the scheduling indicator $J_k^i(\mathbf{H},\boldsymbol{\chi})$ is a deterministic function

of $\mathbf{H}$ and $\boldsymbol{\chi}$. Therefore, the data rate for user k is given by

$$R_k(\mathbf{H}, \boldsymbol{\chi}) = \sum_{i=1}^{M} J_k^i \chi_k \log(1 + \gamma_k^i) \tag{4}$$

where γ_k^i is the SINR of user k on the i-th beam after the Stage II user scheduling.

We define the conditional feedback cost $S(\mathbf{Q})$ and the average feedback cost $\overline{S}$ as follows,

$$S(\mathbf{Q}) = \mathbb{E}\left[\sum_k \chi_k | \mathbf{Q}\right] = \sum_k p_k(\mathbf{Q}), \quad \text{and} \quad \overline{S} = \mathbb{E}\left[S(\mathbf{Q})\right]. \tag{5}$$

In addition, the minimum average feedback cost to achieve the maximum queue stability region $\mathcal{C}$ in the MU-MIMO system is denoted as $\overline{S}^*$.

3) The Feedback Filtering Optimization: The feedback filtering control policy is derived from the Lyapunov technique and is shown to be throughput optimal as follows.

Define $L(\mathbf{Q}) = \sum_k Q_k^2$ as the Lyapunov function. Then the one-step conditional Lyapunov drift $\triangle L(\mathbf{Q}(t))$ is given by,

$$\triangle L(\mathbf{Q}(t)) \quad \triangleq \quad \mathbb{E}\left[L(\mathbf{Q}(t+1) - L(\mathbf{Q}(t)) | \mathbf{Q}(t)\right]. \tag{6}$$

The following lemma establishes the relationship between the Lyapunov drift (6) and the queue stability.

Lemma 1 (Lyapunov drift and the queue stability): Given positive constants V and ϵ, the K queues of the MU-MIMO system $\{Q_1(t), \ldots, Q_K(t)\}$ are stable if the following condition is satisfied,

$$\triangle L(\mathbf{Q}(t)) + VS(\mathbf{Q}(t) | \mathbf{Q}(t)) \leq BK - \epsilon \sum_k Q_k(t) + V\overline{S}^* \tag{7}$$

for all t and all $\mathbf{Q}(t)$. The average queue length satisfies

$$\sum_k \overline{Q}_k \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{\tau=0}^{T-1} \sum_k \mathbb{E}\left[Q_k(\tau)\right] \leq \frac{BK + V\overline{S}^*}{\epsilon} \tag{8}$$

and the average feedback cost satisfies

$$\overline{S} \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{\tau=0}^{T-1} S(\mathbf{Q}(\tau)) \leq \overline{S}^* + BK/V. \tag{9}$$

■

Proof: The proof can be extended from from [18, Lemma 1] by replacing the power cost function with the feedback cost function $S(\mathbf{Q})$ defined in (5). ■

The results in Lemma 1 motivate us to minimize the Lyapunov drift in (7) to achieve the maximum queue stability region. With this insight, we develop our feedback filtering control policy as follows.

Feedback Filtering Control Algorithm (FFCA): Observing the current queue length $\mathbf{Q}(t)$, users feedback their CSI according to the probability vector $\mathbf{p}^*(\mathbf{Q}(t)) = \{p_1^*(\mathbf{Q}(t)), \ldots, p_K^*(\mathbf{Q}(t))\}$, where $\mathbf{p}^*(\mathbf{Q}(t))$ is obtained from the solution of the following optimization problem,

$$\max_{\{p_k\}} \quad \mathbb{E}\Big[\sum_{k=1}^{K} Q_k(t) R_k(\mathbf{H}, \boldsymbol{\chi}) - VS(\mathbf{Q}(t))\Big]. \tag{10}$$

The following theorem justifies the throughput optimality of the feedback filtering policy obtained in (10).

Theorem 1 (Throughput optimality of the FFCA): The feedback control $\mathbf{p}^*(\mathbf{Q})$ given by FFCA achieves the maximum stability region $\mathcal{C}$ in the MU-MIMO system. ■

Proof: Please refer to Appendix A for the proof. ■

From the results in Lemma 1, the parameter V in (10) trades off the average queue length (delay) and the feedback cost. A large parameter V reduces the average feedback cost in (9) but results in a larger average queue length (8). In the next section, we shall derive the FFCA solution $\mathbf{p}^*(\mathbf{Q})$. Note that due to the feedback filtering variable $\boldsymbol{\chi} \in \{0,1\}^K$, we have an exponential complexity (w.r.t. K) to evaluate the expectation in (11). This makes the problem difficult to solve. However, by exploiting specific problem structure, we are still able to find the global optimal solutions to the problem (10).

III. The Queue-Aware User Feedback Filtering Algorithm

In this section, we shall focus on deriving the FFCA solution for feedback filtering problem in (10). Using primal decomposition techniques, problem (10) can be transformed into the following two subproblems

- Inner subproblem:

$$\mathcal{P}_I(S) = \max_{\{p_k\}} \quad \mathbb{E}\left[\sum_{k=1}^K Q_k(t) R_k(\mathbf{H}, \boldsymbol{\chi})\right] \tag{11}$$

$$\text{subject to} \quad 0 \le p_k \le 1, \qquad \forall k = 1, \ldots, K \tag{12}$$

$$\sum_{k=1}^K p_k = S \tag{13}$$

  where S is an auxiliary variable with the physical meaning of the average number of feedback users due to constraint (13).

- Outer subproblem:

$$\mathcal{P}_{II} = \max_{S} \quad \mathcal{P}_I(S) - VS. \tag{14}$$

In the following, we first derive the objective function in (11). We then solve the inner subproblem (11) and outer subproblem (14) separately.

A. The Average Data Rate of the Feedback Users

In this section, we are interested in the expected user data rate $\mathbb{E}[R_k]$ (also denoted as $\overline{R}_k$) in (11). Define $\eta_k(S) \triangleq \mathbb{E}\left[R_k(\mathbf{H}, \boldsymbol{\chi}) \middle| \chi_k = 1, \sum_k \chi_k = S\right]$ as the average data rate for user k, conditioned on S users feedback to the BS (including user k). We characterize $\eta_k(S)$ in the following lemma.

Lemma 2 (Data rate under deterministic feedback): Given the set of feedback users $\mathcal{F}$, where $|\mathcal{F}| = S$, we have for $k \in \mathcal{F}$,

$$\eta_k(S) \le M \int_0^\infty \log(1+x) N f(x) F(x)^{NS-1} dx \triangleq \hat{\eta}_k(S) \tag{15}$$

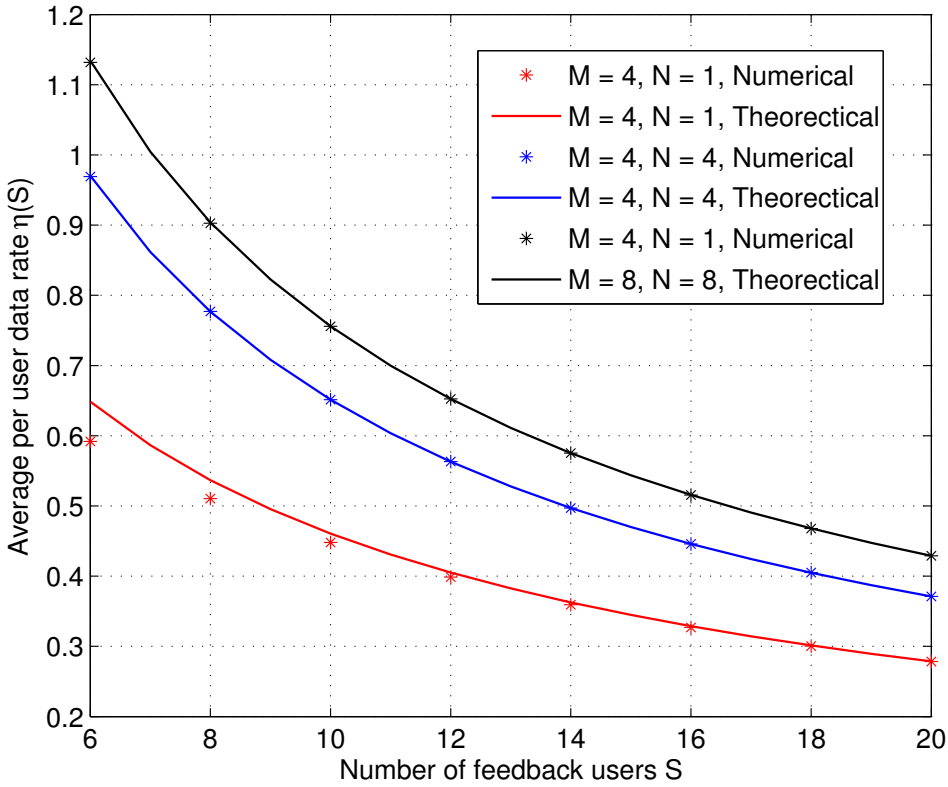

Figure 2. Comparisons between the numerical value $\eta_k(S)$ and its theoretical upper bound $\hat{\eta}_k(S)$ for $k \in \mathcal{F}$. 10^6 channel realizations were run to estimate $\eta_k(S)$. As observed, $\eta_k(S) \approx \hat{\eta}_k(S)$ even for moderate number of users S.

where

$$F(x) = 1 - \frac{e^{-x/P}}{(1+x)^{M-1}}. \tag{16}$$

is the cumulative distribution function (CDF) of $\text{SINR}_{k,n}^i$ in (2) and $f(x)$ is the corresponding probability distribution function (PDF).

Moreover, the upper bound is tight as

$$\triangle\eta = |\hat{\eta}_k(S) - \eta_k(S)| \leq \left(1 - \frac{e^{-1/P}}{2^{M-1}}\right)^{NS}.$$

■

Proof: Please refer to Appendix B for the proof. ■

Fig. 2 illustrates the comparisons between $\eta_k(S)$ and $\hat{\eta}_k(S)$ for different number of users S. As observed, $\eta_k(S) \approx \hat{\eta}_k(S)$ even for moderate number of users S.

B. Solution of the FFCA

1) Solution to the inner subproblem: In the following, we shall utilize Lemma 2 to solve for the inner problem. Let $\Pi = \{\pi(1), \ldots, \pi(K)\}$ be a permutation of $\mathbf{Q}$ such that $Q_{\pi(1)} \geq Q_{\pi(2)} \geq \cdots \geq$

$Q_{\pi(K)}$. Note that the distribution of the binary random variable χ_k is specified by p_k, and we have $\mathbb{E}[\chi_k] = p_k$. Given the average feedback amount $\mathbb{E}[\sum \chi_k] = \sum_k p_k = S$, the FFCA solution $\mathbf{p}$ of the problem (11) is summarized in the following theorem.

Theorem 2 (The optimal user feedback probability $\{p_k\}$): The FFCA user feedback probability $\{p_k\}$ to solve (11) is given by

$$p_{\pi(k)} = 1, \qquad 1 \le k \le \lfloor S \rfloor \tag{17}$$

$$p_{\pi(k_0)} = S - \lfloor S \rfloor, \quad k_0 = \lfloor S \rfloor + 1 \tag{18}$$

$$p_{\pi(k)} = 0, \qquad \text{otherwise.} \tag{19}$$

and the average data rate for user $k \le k_0$ is

$$\begin{aligned} \overline{R}_k &= \eta_k(\lfloor S \rfloor)\left(1 - (S - \lfloor S \rfloor)\right) + \eta_k(\lfloor S \rfloor + 1)\left(S - \lfloor S \rfloor\right) \\ \overline{R}_{k_0} &= \eta_{k_0}(\lfloor S \rfloor + 1)\left(S - \lfloor S \rfloor\right). \end{aligned}$$

■

Proof: Please refer to Appendix C for the proof. ■

The above result shows that given the constraint on the average number of feedback users S, the best strategy is to let the users with the S largest queues to feedback, while keeping other users inactive.

2) Solution to the outer subproblem: Based on the solution of the inner subproblem (11), we are now ready to solve the outer subproblem (14) and determine the average feedback cost S^*. Using the results for $\overline{R}_k$ in Theorem 2, we obtain the objective function of the outer subproblem (14) as follows

$$\begin{aligned} U(S) &= \mathbb{E}\left[\sum_{k=1}^{\lfloor S \rfloor} Q_{\pi(k)} R_{\pi(k)} | \chi_{\pi(k_0)} = 0\right]\left(1 - p_{\pi(k_0)}\right) \\ &\quad + \mathbb{E}\left[\sum_{k=1}^{\lfloor S \rfloor + 1} Q_{\pi(k)} R_{\pi(k)} | \chi_{\pi(k_0)} = 1\right] p_{\pi(k_0)} - VS \\ &= \sum_{k=1}^{\lfloor S \rfloor} Q_{\pi(k)} \eta_{\pi(k)}(\lfloor S \rfloor)\left[1 - (S - \lfloor S \rfloor)\right] \\ &\quad + \sum_{k=1}^{\lfloor S \rfloor + 1} Q_{\pi(k)} \eta_{\pi(k)}(\lfloor S \rfloor + 1)\left(S - \lfloor S \rfloor\right) - VS. \end{aligned}$$

Note that $U(S)$ is a continuous function on S. The following result shows that it suffices to consider only integer value of S.

Lemma 3 (Property of the outer subproblem): The solution S^* to the outer subproblem (14) is an integer. ■

Algorithm 1 Proposed algorithm to find user feedback filtering policy on stage I.

1) Initialization: $S = \lfloor \frac{K}{2} \rfloor$. $S_{\min} = 1$, $S_{\max} = K$.
2) Evaluate the condition in (21). If $U(S^*) \geq U(S^* - 1)$, then $S_{\min} = S$. Otherwise, $S_{\max} = S$.
3) Repeat 2) by setting $S = \lfloor (S_{\min} + S_{\max})/2 \rfloor$, until $S_{\max} - S_{\min} \leq 1$.
4) Find the optimal user feedback probability vector $\mathbf{p}$ according to (17) in Theorem 2, by setting $S = S^*$ found from the above. The user feedback filtering policy on stage I is thus determined.

Proof: Please refer to Appendix D for the proof. ■

With Lemma 3 the outer subproblem in (14) becomes

$$\begin{aligned} \max_{S} \quad & U(S) = \textstyle\sum_{k=1}^{S} Q_{\pi(k)} \eta_{\pi(k)}(S) - VS \\ \text{subject to} \quad & S = \{1, \ldots, K\}. \end{aligned} \tag{20}$$

An intuitive observation of the outer subproblem (20) is that, while the term $\sum_{k=1}^{S} Q_{\pi(k)}$ in (20) is increasing with S, and the terms $\eta(S)$ and $-VS$ are decreasing, the objective $U(S)$ should first increase and then decrease after it reaches $U(S^*)$. This motivates us to use a bisection algorithm (summarized in Algorithm 1) to solve the outer subproblem (20), which takes maximum $\log_2(K)$ steps to find S^*. The following theorem guarantees that the bisection algorithm finds the global maximum of the outer subproblem (20).

Theorem 3 (Global optimal solution to (10)): The global optimal solution to (10) is uniquely determined by the following conditions

$$U(S^*) \geq U(S^* + 1) \text{ and } U(S^*) \geq U(S^* - 1) \tag{21}$$

where $S^* \in \{1, \ldots, K\}$. ■

Proof: Please refer to Appendix E for the proof. ■

Using with Theorem 2 and 3, the two timescale user scheduling algorithm can be summarized as follows. We first determine the optimal user feedback amount S^* by solving (20) using Algorithm 1. We then choose S^* users who have the longest queues among all the K users to be eligible to feedback the BS, and this feedback filtering decision $\{p_k^*(\mathbf{Q})\}$ in (17) is broadcasted to the network. As a result, the users feedback their effective SINRs based on $\{p_k^*(\mathbf{Q})\}$ and the BS schedules the users based on their SINR as described in the stage II policy.

IV. Large Deviation Delay Analysis for the Worst Case User

In this section, we will study the queueing delay performance of the proposed solution and illustrate the gain of having queue-aware policy. We are interested in the steady state distribution of the worst

case queueing performance, i.e.,

$$\lim_{t\to\infty} \Pr(\max_{1\le k\le K} Q_k(t) > B)$$

where B is the buffer size. We denote $Q_{\max}(t) = \max_k Q_k(t)$ as the maximum queue length process and $Q_{\max}(\infty)$ as the steady state of the $Q_{\max}(t)$. To overcome the technical challenges associated with delay analysis of MU-MIMO system, we consider the large deviation approach [19]. Specifically, we focus on the asymptotic overflow probability for the maximum queue $Q_{\max}(\infty)$ over a large buffer size B, which is captured by the large deviation decay rate of the tail probability of $Q_{\max}(\infty)$. In the next section, we shall introduce the decay rate function for $Q_{\max}(\infty)$.

A. Large Deviation Decay Rate for $Q_{\max}(\infty)$ Using Sample Path Analysis

The large deviation decay rate function I^* for the tail probability of $Q_{\max}(\infty)$ is defined as

$$I^* \triangleq \lim_{B\to\infty} -\frac{1}{B}\log \Pr\left(Q_{\max}(\infty) > B\right). \tag{22}$$

Note that, with the notion of the large deviation rate function, the queue overflow probability can be written as

$$\Pr(Q_{\max}(\infty) > B) = e^{-I^*B+o(B)} \tag{23}$$

where the component I^* controls how fast the queue overflow probability drops when the buffer size B grows. A larger decay rate I^* corresponds to a better performance of the scheduling algorithm in the sense of reducing the worst case delay $Q_{\max}$ in the system. In the following, we shall find the decay rate function I^*.

Consider a scaled sample path $q^B_{\max}(t) = \frac{1}{B}Q_{\max}(\lfloor Bt\rfloor)$, which starts from $q^B_{\max}(0) = 0$ and reaches $q^B_{\max}(T_s) = 1$, for some T_s. Note that with the scaling, we have $\Pr(Q_{\max}(\infty) > B) = \Pr\left(q^B_{\max}(\infty) > 1\right)$. Let $w(t)$ be a continuous sample path following $q^B_{\max}(t)$, as $w(t) \approx q^B_{\max}(t)$. Computing the decay rate I^* corresponds to finding a "most likely" path $w(t)$ that overflows at $w(T_s) = 1$. Using the *large deviation principle* [19], the decay rate function I^* can be found as follows

$$I^* = \inf\left\{\int_0^{T_s} l(w(\tau), w'(\tau))d\tau : w(0) = 0, w(T_s) = 1, T_s > 0\right\}$$

where $l(w(t), w'(t))$ defined in (38) is the *local rate* function [19] following the path $w(t)$ (see Appendix G).

Solving the above variational calculus problem we obtain the results as follows. Denote $\mu_k(x) = \frac{\overline{R}_k(x)}{p_k(x)L}$ for $k \in \mathcal{F}$, where $x = q^B_{\max}(t)$ for $0 \le t \le T_s$. Note that from (15), $\mu_k(x)$ is independent of k and thus we write $\mu(x) = \mu_k(x)$. Consider the arrivals to all the users are i.i.d. with mean $\mathbb{E}[A_k] = \lambda$ and logarithm moment generating function $g_A(\theta) = \log \Lambda_{A,k}(\theta)$. The following theorem summarizes the large deviation decay rate for $Q_{\max}(\infty)$ under the proposed two timescale algorithm.

Theorem 4 (The large deviation decay rate for $Q_{\max}(\infty)$): Suppose $\lambda/\mu(x) < 1$ for all x in $[0,1]$. Then the large deviation decay rate for $Q_{\max}(\infty)$ is given by

$$I^* = \int_0^1 \theta^*(x)dx \tag{24}$$

where $\theta^*(x)$ solves

$$g(x,\theta) = g_A(\theta) + \mu(x)\left(e^{-\theta} - 1\right) = 0.$$

■

Proof: Please refer to Appendix G for the proof. ■

The result in Theorem 4 gives the rate function to evaluate the exponential decay rate of the overflow probability $\Pr(Q_{\max}(\infty) > B)$ under the proposed two timescale user scheduling algorithm. Based on this result, we shall derive more insights in the later sections.

B. Approximation of the Average User Data Rate $\overline{R}_k$

From the expression of the rate function I^* in (24), it is still hard to understand how the delay performance relates to the system parameters. To further analyze I^*, we need a closed form expression for the average data rate $\overline{R}_k$ for each user. In this section, we shall derive an asymptotically accurate approximation for $\overline{R}_k$.

Lemma 4 (Asymptotic analysis of $\overline{R}_k$): The average data rate $\overline{R}_k$ for $k \in \mathcal{F}$ has the following property,

$$\lim_{S\to\infty} \frac{\overline{R}_k}{p_k \frac{M}{S} \log\left(P \log NS\right)} = 1. \tag{25}$$

■

Proof: This is a direct result of [2, Theorem 1] by considering $n = NS$ users each with single receive antenna. ■

Lemma 4 shows that we can use $\overline{R}_k \approx p_k \frac{M}{S} \log\left(P \log NS\right)$ as an asymptotically accurate approximation of the average data rate, when S is large. Using the approximated $\overline{R}_k$, we find an upper bound for S^* in the following.

Lemma 5 (Upper bound of S^*): The upper bound of S^* which solves (36) is given by

$$S^*(Q_{\max}; K) \leq \min\left\{\frac{1}{N} e^{W(c_1)}, K\right\} \tag{26}$$

where $c_1 = \frac{MNQ_{\max}}{V}$, and $W(x)$ is the Lambert W function [20] defined as $W(x)e^{W(x)} = x$. The equality holds when $Q_{\pi(k)} \equiv Q_{\max}$ for all k. ■

Proof: Please refer to Appendix F for the proof. ■

Remark 1 (Interpretation of S^*): The results show that, when $Q_{\max}$ is large, it is better to have more user feedback to boost up the system throughput. On the other hand, when $Q_{\max}$ is small, we can have less user feedback and give higher priorities to the urgent users. Note that according to

Lemma 2, the average data rate $\overline{R}_k$ is a decreasing function of S. Thus the upper bound of S^* gives a lower bound of $\overline{R}_k$, i.e.,

$$\overline{R}_k(Q_{\max}) \geq p_k(\hat{S}^*(Q_{\max}))M\frac{\log\left(P\log N\hat{S}^*(Q_{\max})\right)}{\hat{S}^*(Q_{\max})} \tag{27}$$

where $\hat{S}^*(Q_{\max}) \triangleq \frac{1}{N}\exp\left(W(\frac{MNQ_{\max}}{V})\right)$ and $p_k(\hat{S}^*(Q_{\max}))$ is given by Theorem 2. Using this, we can find a lower bound expression of decay rate function I^* in the next section.

C. Asymptotic Analysis and Comparisons with the CSI-only User Scheduling

In this section, we shall derive some asymptotic results for I^* to get a better insight in the behavior of the worst case delay. As a comparison, the delay performance under a CSI-only baseline policy will also be derived.

The CSI-only baseline algorithm assumes that each user k feeds back the SINR for the $i^*(k,n)$-th beam on each antenna n, where $i^*(k,n) = \arg\max_{1\leq i\leq M} \text{SINR}_{k,n}^i$. Then for each beam i, the BS schedules the user who has the highest SINR on beam i. The CSI-only baseline scheme corresponds to a special case of the proposed two timescale user scheduling by setting $\chi_k \equiv 1$ for all k in stage I.

In order to obtain more design insights from the results in Theorem 4, we shall consider a special case where the arrivals A_k follow Poisson distributions. Specifically, the MGF of A_k is given by $\Lambda_{A,k}(\theta) = e^{\lambda(e^\theta-1)}$. We have the following results.

Corollary 1 (Decay rate for the CSI-only algorithm): Let $\mu_b = \frac{M\log(P\log NK)}{KL}$ and $\lambda_T = \lambda K$. For Poisson arrivals with $\lambda < \mu_b$, the large deviation decay rate of $Q_{\max}(\infty)$ under the CSI-only baseline algorithm can be expressed as

$$I^*_{\text{baseline}} = \log\frac{M\log\left(P\log NK\right)}{\lambda_T L}. \tag{28}$$

■

Proof: Please refer to Appendix H for the proof. ■

Remark 2 (Interpretation of the results): We can observe that, under a fixed total arrival rate λ_T, the CSI-only baseline algorithm has a decay rate $I^* = \mathcal{O}(\log\log\log K)$ due to the standard multi-user diversity gain.

Similarly, we obtain the following results for the large deviation decay rate of $Q_{\max}(\infty)$ under the proposed two timescale user scheduling algorithm.

Corollary 2 (Decay rate for the proposed algorithm): Let $\mu_0 = \inf_{x\in[0,1]}\mu_p(x)$ and $\lambda_T = \lambda K$, where $\mu_p(x) = \frac{M\log\left(P\log N\hat{S}^*(x)\right)}{L\hat{S}^*(x)}$. Under Poisson arrivals with $\lambda < \mu_0$, the large deviation decay rate of $Q_{\max}(\infty)$ under the two timescale user scheduling algorithm can be expressed as

$$I^*_{\text{prop}} \geq (1-\epsilon_s)\log K + \log\frac{M}{\lambda_T L} + \epsilon_s\log R_0 + C \tag{29}$$

where $C = \int_{\epsilon}^{1} \left\{ \log \left[N \log \left(PW \left(\frac{MNx}{V} \right) \right) \right] - W \left(\frac{MNx}{V} \right) \right\} dx$, $R_0 = \int_0^1 \log\left(1 + Px\right) dF(x)$, and $\epsilon > 0$ is a small constant. ■

Proof: Please refer to Appendix I for the proof. ■

Based on the results in Corollary 1 and 2 we conclude the following for the CSI-only user scheduling algorithm and the proposed two timescale algorithm.

- Large deviation decay rates $I^*_{\text{prop}} \gg I^*_{\text{baseline}}$, when the number of users K grows large. This demonstrates that it is important to utilize the queue information in the user scheduling algorithm to minimize the worst case delay.
- In addition, both of the schemes benefit from the multiuser diversity. The decay rate increases when the number of users increases, and the rate I^*_{prop} increases faster than the baseline.
- Furthermore, both of the schemes benefit from the MU-MIMO channel. It is demonstrated that, when increasing the number of data streams M and the receive antennas N, the large deviation decay rates I^*_{prop} and I^*_{baseline} both increase as $\mathcal{O}(\log M \log\log N)$.

In summary, by carefully exploiting the queue information in the stage I feedback filtering, the proposed MU-MIMO algorithm has significant delay performance gain compared with conventional CSI-only schemes.

V. Numerical Results

In this section, we simulation the queueing delay performance of the proposed two timescale user scheduling algorithm. We consider a MU-MIMO system with K users, and packets arrive to the queue of each user according to a Poisson distribution with rate $\lambda = \lambda_T/K$, where the total arrival rate is $\lambda_T = 7500$ packets/second. Each packet has $L = 8000$ bits. The system bandwidth is 10 MHz and the SNR is 10 dB. The number of transmit and receive antennas are $M = 4$ and $N = 2$, respectively. The scheduling time slot is $\tau = 1$ ms and the simulation is run over $T_{tot} = 100$ seconds. We compare the performance of proposed algorithm against the following reference baselines.

- **Baseline 1: CSI-only user scheduling (CSIO)** [6]. At each time slot, all the users feedback the CSI to the BS, and the BS schedules a set of users who respectively have the highest SINR on each beam (see Section IV-C).
- **Baseline 2: CSI-only user scheduling with limited feedback (CSIO-LF)** [6]. The scheme is similar to baseline 1 except that the user feeds back to the BS only when its SINR exceeds a threshold $t_{SINR} = 2$ dB.
- **Baseline 3: Proportional fair user scheduling (PFS)** [1]. At each time slot, all the users feedback the CSI to the BS, and the BS transmits data to the users using proportional fair scheduling with window size $t_w = 100$ ms.

- **Baseline 4: Max weighted queue user scheduling (MWQ)** [13]. At each time slot, all the users feedback their CSI to the BS, and the BS selects a set of users so that the instantaneous queue-weighted sum rate $\sum Q_k R_k$ is maximized.

Note that the associated user scheduling problem in baseline 4 has much higher complexity for user scheduling and feedback from all the users are required. Hence, baseline 4 serves for performance benchmarking purpose only.

A. Queueing Performance and Feedback Comparisons

Fig. 3 shows the overflow probability for the worst case queue $\Pr\left(Q_{\max}(\infty) > B\right)$ versus the buffer size B. The number of users is $K = 40$. The feedback policy $\boldsymbol{\chi}$ updates on every $T = 1, 5, 10$ time slots. The proposed scheme significantly outperforms over baselines 1 - 3. It also has similar performance as baseline 4. Fig. 4 demonstrates the average feedback amount $\overline{S}$ (defined as the average number of users feedback to the BS at each time slot) versus the number of users K. The feedback amount of the proposed scheme is less than those of all the baselines. Note that although baseline 4 has a smaller worst case queue, it requires all the users feedback to the BS.

B. Large Deviation Decay Rate for a Large Number of Users

Fig. 5 the large deviation decay rate over the number of users. The decay rates I^* in (22) are evaluated at buffer size $B_{0.05}$, where overflow probability $\Pr\left(Q_{\max}(\infty) > B_{0.05}\right) = 0.05$. The decay rate for the proposed scheme grows much faster than those of baselines 1 - 3 with the number of users K. This is consistent with the result in Corollary 2.

VI. Conclusions

In this paper, we proposed a novel two timescale delay-aware user scheduling algorithm for the MU-MIMO system. The policy consists of a queue-aware mobile-driven feedback filtering stage and a dynamic SINR-based user scheduling stage. The queue-aware feedback filtering control algorithm in stage I was derived through solving an optimization problem. Under the proposed two timescale user scheduling algorithm, we also evaluated the queueing delay performance for the worst case user using the sample path large deviation analysis. The large deviation decay rate for the proposed algorithm, scaled as $\mathcal{O}\left(\log K\right)$, was shown to be much larger than a CSI-only user scheduling algorithm, which means that the proposed scheme performs better in reducing the worst case delay. The numerical results demonstrated a significant performances gain over the CSI-only algorithm and a huge feedback reduction over the MWQ algorithm.

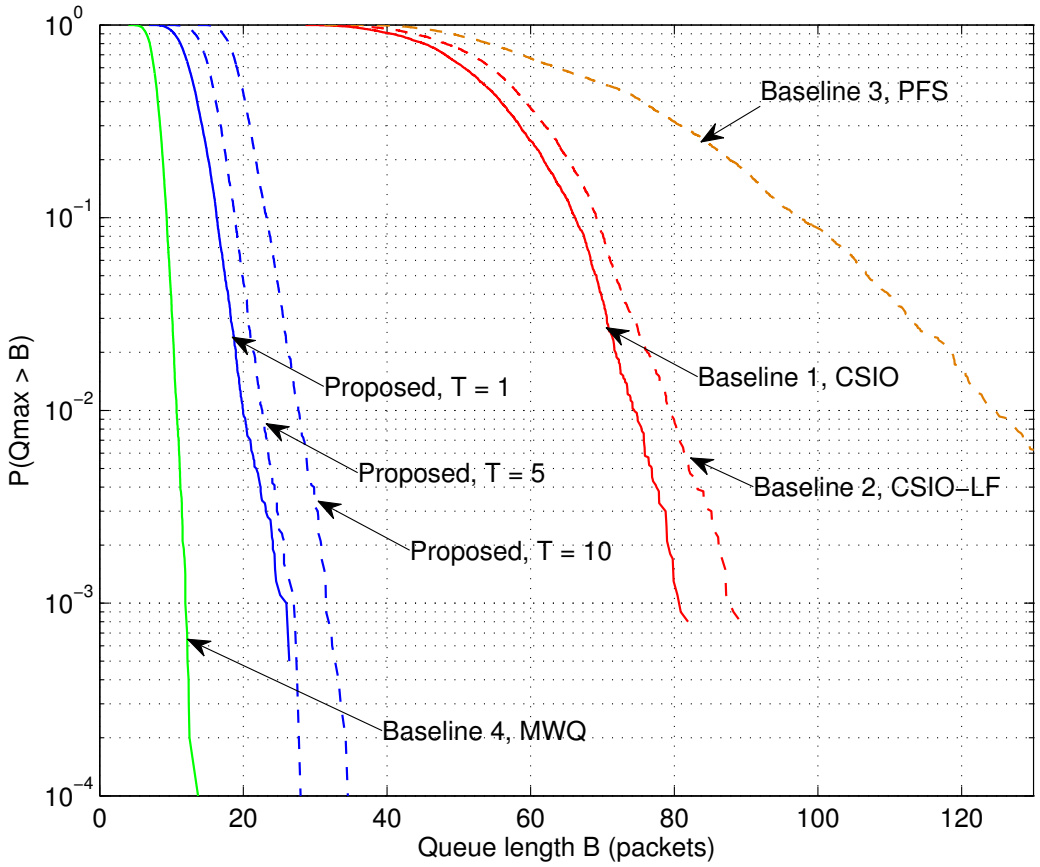

Figure 3. The overflow probability for the worst case queue $\Pr\left(Q_{\max}(\infty) > B\right)$ versus the buffer size B. The number of users is $K = 40$. The feedback policy $\boldsymbol{\chi}$ in stage I updates on every $T = 1, 5, 10$ time slots. The proposed scheme significantly outperforms over baselines 1 - 3. It also performs closely to baseline 4.

APPENDIX A
THROUGHPUT OPTIMAL PROPERTY OF THE FFCA POLICY

In this section, we prove that the feedback control policy (10) can achieve the maximum stability region.

Consider the queue dynamic in (3). By squaring the equation on both sides and using the property $[\max\{0, x\}]^2 \leq x^2$, we obtain $\forall k$,

$$Q_k^2(t+1) \leq Q_k^2(t) + \mu_k^2(t) - 2Q_k(t)(\mu_k(t) - A_k(t)) + A_k^2(t) \tag{30}$$

where we simplify the notation by writing $\mu_k(\mathbf{Q}(t), \mathcal{H}(t))$ as $\mu_k(t)$. Following the definition of conditional Lyapunov drift $\triangle L(\mathbf{Q}(t))$ in (6), taking conditional expectations and summing over all

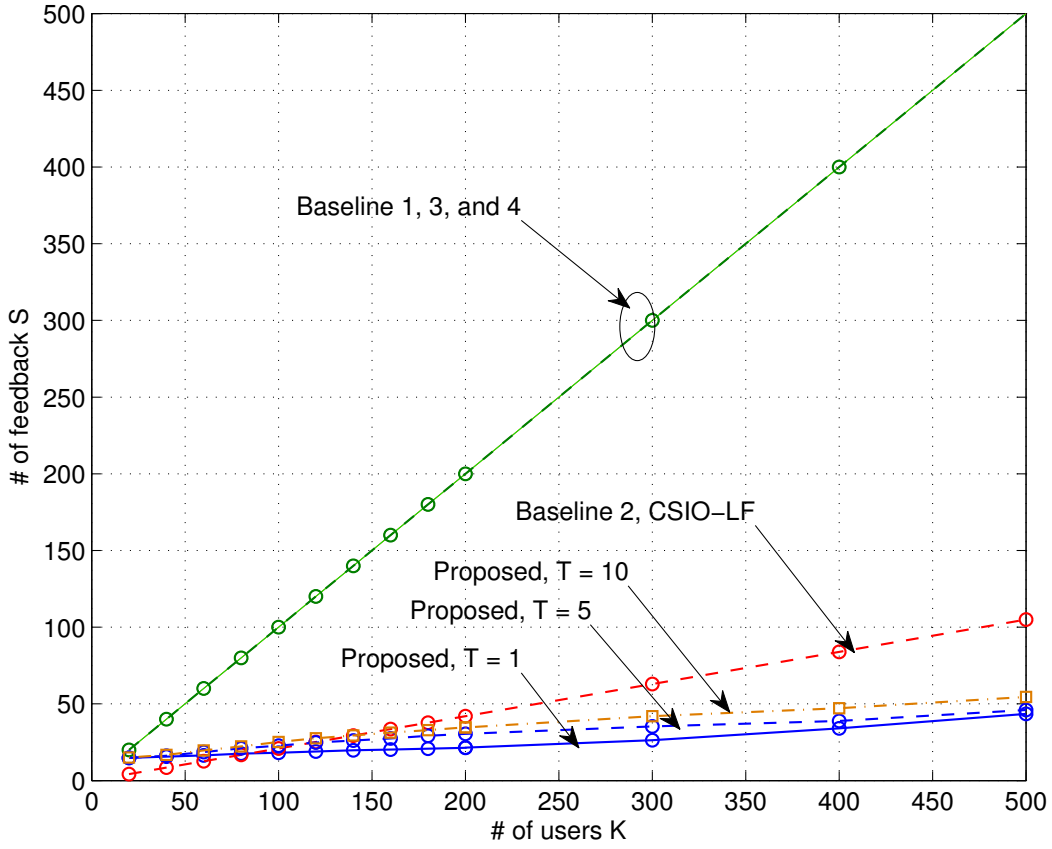

Figure 4. The average feedback amount $\overline{S}$ versus the number of users K. The feedback threshold of baseline 2 is $t_{SINR} = 2$ dB. The feedback amount of the proposed scheme is much less than those of all the baselines. Note that although baseline 4 (MWQ) has a smaller worst case queue, it requires all the users feedback to the BS.

k inequalities in (30) yields

$$\begin{aligned} \triangle L(\mathbf{Q}(t)) \quad \leq \quad & \mathbb{E}\left[\sum_k \mu_k^2(t) + A_k^2(t)|\mathbf{Q}(t)\right] \\ & -2\sum_k Q_k(t)\mathbb{E}\left[\mu_k(t) - A_k(t)|\mathbf{Q}(t)\right]. \end{aligned} \tag{31}$$

Denote positive constants $\overline{\mu}_{\max}^2$ and $\overline{\lambda}_{\max}^2$ such that

$$\mathbb{E}\left[\mu_k^2(t)|\mathbf{Q}(t)\right] \leq \overline{\mu}_{\max}^2 \text{ and } \mathbb{E}\left[A_k^2(t)|\mathbf{Q}(t)\right] \leq \overline{\lambda}_{\max}^2.$$

Let $B = \overline{\mu}_{\max}^2 + \overline{\lambda}_{\max}^2$. The drift (31) is bounded by

$$\triangle L(\mathbf{Q}(t)) \leq BK - 2\sum_k Q_k(t)\left\{\mathbb{E}\left[\mu_k(t)|\mathbf{Q}(t)\right] - \lambda_k\right\}. \tag{32}$$

Suppose now that the arrival $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ is strictly interior to the stability region $\mathcal{C}$ (Definition 2) such that $\boldsymbol{\lambda} + \epsilon\mathbf{1} \in \mathcal{C}$, for $\epsilon > 0$. Since channel states are i.i.d. over time slots, using the result

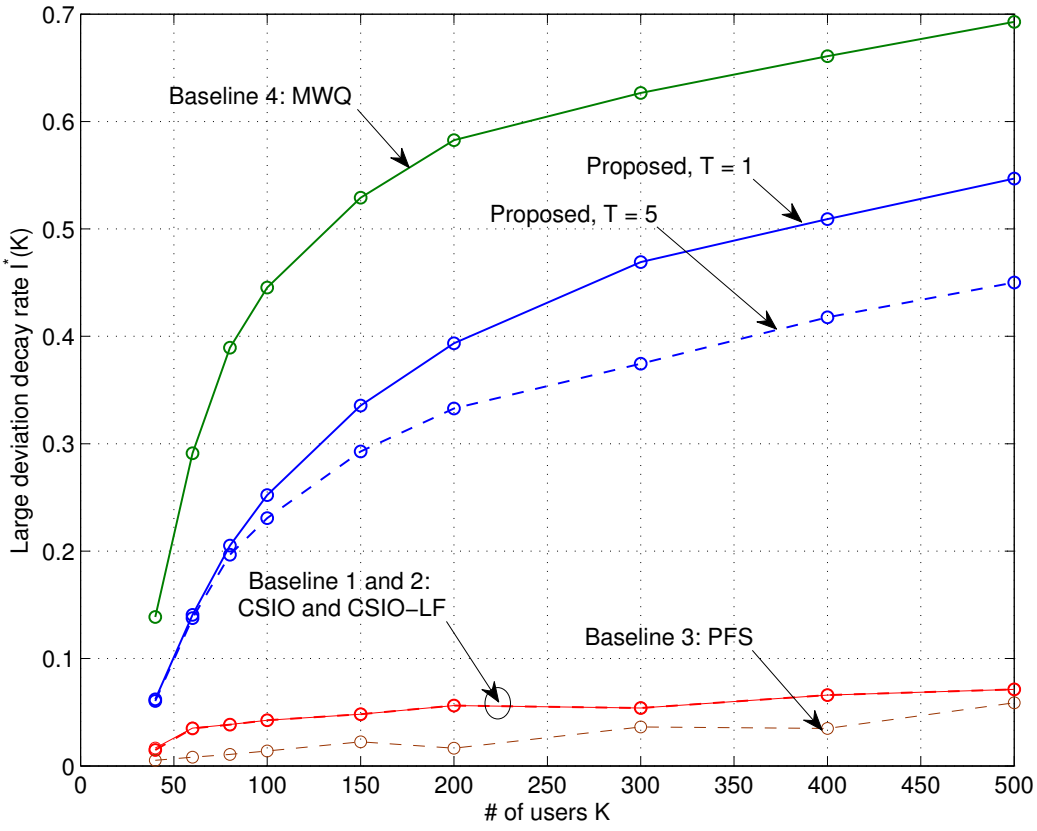

Figure 5. The large deviation decay rate over the number of users. The decay rates I^* in (22) are evaluated at buffer size $B_{0.05}$, where overflow probability $\Pr(Q_{\max}(\infty) > B_{0.05}) = 0.05$. The decay rate for the proposed scheme grows much faster than that of baselines 1 - 3 with the number of users K. Note that although baseline 4 performs the best, it requires all the users feedback to the BS.

in [18, Corollary 1], it follows that there exists a stationary randomized feedback control policy that schedules user to feedback independent of queue $\mathbf{Q}(t)$ and yields

$$\begin{aligned}\mathbb{E}\left[\mu_k(t)|\mathbf{Q}(t)\right] = \mathbb{E}\left[R_k(t)\right] &\geq \lambda_k + \epsilon, \qquad \forall k \\ \mathbb{E}\left[S(\mathbf{Q}(t)|\mathbf{Q}(t)\right] &= \overline{S}(\epsilon).\end{aligned}$$

Because the stationary policy is simply a particular feedback policy and note that the EECA maximizes the term $\sum_k \mathbb{E}\left[Q_k(t)R_k(t)\right]$, the right hand side of (32) under FFCA is less than or equal to the resulting value under the stationary policy. Therefore, we have

$$\triangle L(\mathbf{Q}(t)) + VS(\mathbf{Q}(t)|\mathbf{Q}(t)) \leq BK - 2\epsilon\sum_k Q_k(t) + V\overline{S}(\epsilon).$$

Notice that $\overline{S}(\epsilon) \leq K$, which is the maximum feedback cost by the definition of the cost function in (7). Using the results in Lemma 1, it follows that $\sum_k Q_k(t) \leq \frac{BK+V\overline{S}(\epsilon)}{2\epsilon} \leq \frac{BK+VK}{2\epsilon} < \infty$, which

proves that the FFCA policy stabilizes all the queues.

APPENDIX B
DATA RATE FOR THE USER WITH MAXIMUM SINR

From the effective SINR expression in (2), as ϕ_i are unitary vectors, $|H_k^{(n)}\phi_i|^2$ are i.i.d. over i with chi-square distribution with degrees of freedom 2. Consequently, the term $\sum_{j:j\neq i}\left|H_k^{(n)}\phi_j\right|^2$ is chi-square distributed with degrees of freedom $2M-2$. Thus, the PDF $f(x)$ and CDF $F(x)$ of $\text{SINR}_{k,n}^i$ are given by [2]

$$f(x)=\frac{e^{-x/P}}{(1+x)^M}\left(\frac{1}{P}(1+x)+M-1\right)$$

and

$$\begin{aligned}F(x) &= \int_0^x \frac{e^{-x/P}}{(1+x)^M}\left(\frac{1}{P}(1+x)+M-1\right)dx\\ &= 1-\frac{e^{-x/P}}{(1+x)^{M-1}},\end{aligned}$$

respectively. Thus, for a particular user $k\in\mathcal{F}$, as $\text{SINR}_{k,n}^i$ are i.i.d. over different users k and antennas n, the probability that user k has the largest SINR on the i-th beam and the n-th antenna is give by $1/NS$. The corresponding CDF of the maximum SINR is

$$\begin{aligned}P(\max_{k\in\mathcal{F},1\leq n\leq N}\text{SINR}_{k,n}^i\leq x) &= \Pi_{k,n}\text{Pr}\left(\text{SINR}_{k,n}^i\leq x\right) &(33)\\ &= (F(x))^{NS} &(34)\end{aligned}$$

and hence, the data rate can be given by

$$\begin{aligned}R &= \mathbb{E}_{\mathbf{H}}\left[\log(1+\max_{k\in\mathcal{F},1\leq n\leq N}\text{SINR}_{k,n}^i)\right]\\ &= \int_0^\infty \log(1+x)d(F(x))^{NS}\\ &= \int_0^\infty \log(1+x)NSf(x)F(x)^{NS-1}dx.\end{aligned}$$

As each user equips with N antennas, the average data rate for user $k\in\mathcal{F}$, given $|\mathcal{F}|=S$ is

$$\begin{aligned}\eta(S) &\leq \sum_{n=1}^N\sum_{i=1}^M \text{Pr}\left(\text{SINR}_{k,n}^i=\max_{k_0\in\mathcal{F},1\leq n\leq N}\text{SINR}_{k_0,n}^i\right)R\\ &= NM\frac{1}{NS}R\\ &= M\int_0^\infty \log(1+x)Nf(x)F(x)^{NS-1}dx\end{aligned}$$

where this is an upper bound since there is a small probability that a user has maximum SINRs for more than 2 beams on one antenna. As the feedback policy only allows the user to pick up one beam for each antenna to feedback, hence decreases the throughput.

However, the probability that a user has maximum SINRs for more than 2 beams on one antenna is very small, which can be shown in the following lemma.

Lemma 6: Given $\max_{k\in\mathcal{F},1\le n\le N}\text{SINR}_{k,n}^{i}\ge 1$, $\forall i=1\dots M$, it is impossible for a user to have maximum SINRs for more than two beams on one antenna, i.e., for $(k^*,n^*)=\arg\max_{k\in\mathcal{F},1\le n\le N}\text{SINR}_{k,n}^{i}$, we have $\text{SINR}_{k^*,n^*}^{i}=\max_{1\le j\le M}\text{SINR}_{k^*,n^*}^{j}$, $\forall i$. ■

Proof: Given $\text{SINR}_{k^*,n^*}^{i}=\max_{k\in\mathcal{F},1\le n\le N}\text{SINR}_{k,n}^{i}\ge 1$, we have

$$|H_{k^*}^{(n^*)}\phi_i|^2\ge 1/P+\sum_{j\neq i}|H_{k^*}^{(n^*)}\phi_j|^2\ge|H_{k^*}^{(n^*)}\phi_j|^2.$$

Therefore,

$$\text{SINR}_{k^*,n^*}^{j}=\frac{|H_{k^*}^{(n^*)}\phi_j|^2}{1/P+\sum_{r\neq j}|H_{k^*}^{(n^*)}\phi_r|^2}<\frac{|H_{k^*}^{(n^*)}\phi_j|^2}{|H_{k^*}^{(n^*)}\phi_i|^2}\le 1$$

which shows that $\text{SINR}_{k^*,n^*}^{i}$ is the maximum for user k^* on antenna n^* over all the M beams. ■

In the following, we evaluate the gap $\triangle\eta=\hat{\eta}(S)-\eta(S)$. Consider the maximum SINRs for beam i and j are on antenna n^* of user k^*. Assume $\text{SINR}_{k^*,n^*}^{i}\ge\text{SINR}_{k^*,n^*}^{j}$. Thus we must have $\text{SINR}_{k^*,n^*}^{i}<1$ from the above lemma. Therefore, the lost of not reporting $\text{SINR}_{k^*,n^*}^{j}$ due to the stage I policy is bounded by

$$\begin{aligned}\triangle\eta &\le M\Pr\left(\max_{k\in\mathcal{F},1\le n\le N}\text{SINR}_{k,n}^{i}<1\right)\cdot\log(1+\text{SINR}_{k^*,n^*}^{j})\\ &= M\Pr\left(\text{SINR}_{k,n}^{1}<\right)^{NS}\log 2\\ &= \left(1-\frac{e^{-1/P}}{2^{M-1}}\right)^{NS}.\end{aligned}$$

APPENDIX C
POOF OF THEOREM 2

According to Lemma 2, the objective in (11) can be written as $f(\mathbf{p})=\sum Q_k(t)\mathbb{E}\left[\chi_k\eta(s)\right]$, where $\mathbf{p}=\{p_1,\dots,p_K\}$ determines the distributions of χ_k and s. We attempt this problem by considering the following two cases.

Case 1: We consider $p_k=1$ for some $k\in\mathcal{K}_1\subset\mathcal{K}$, with $|\mathcal{K}_1|=S_1<S$, and $p_k<1$ for some $k\in\mathcal{K}_0=\mathcal{K}\backslash\mathcal{K}_1$, where $\mathcal{K}$ is the set for all users. We thus apply Poisson approximation to determine the distribution of s, where $s=S_1+s'$ and s' satisfies Poisson distribution with mean $\nu=S-S_1$, and thus, we have $\sum_{k\in\mathcal{K}\backslash\mathcal{K}_1}p_k=\nu$.

Using the Poisson approximation for the distribution of s decouples χ_k and s. Hence we have $\mathbb{E}\left[\chi_k\eta(s)\right]\approx p_k\overline{\eta}(s)$, where $\overline{\eta}(s)=\mathbb{E}\left[\eta(s)\right]$ is independent of p_k. Therefore, the inner subproblem becomes a linear programming problem as

$$\begin{aligned}\max_{\{p_k\}}\quad & \textstyle\sum_{k=1}^{K}p_kQ_k\overline{\eta}(s)\\ \text{subject to}\quad & \text{constraints (12)}-\text{(13)}\end{aligned}$$

where the solution is given by $p_{\pi(k)} = 1,\ 1 \leq k \leq \lfloor S \rfloor$, $p_{\pi(k_0)} = S - \lfloor S \rfloor,\ k_0 = \lfloor S \rfloor + 1$, and $p_{\pi(k)} = 0$, otherwise. Here, the permutation $\Pi = \{\pi(k)\}$ is such that $Q_{\pi(1)} \geq \cdots \geq Q_{\pi(K)}$. However, the solution violates the assumption of $p_k < 1$ for some $k \notin \mathcal{K}_1$. This leads to the second case as follows.

Case 2: We consider $p_k = 1$ for some $k \in \mathcal{K}_1 \subset \mathcal{K}$, with $|\mathcal{K}_1| = \lfloor S \rfloor \triangleq S_1$, and $p_k = S - \lfloor S \rfloor$ for some $k = k_0$. The inner subproblem then becomes

$$\begin{aligned} \max_{\{p_k\}, k_0} \quad & \textstyle\sum_{k \neq k_0} p_k Q_k \eta(S_1)\,(1 - p_{k_0}) \\ & + \left(\textstyle\sum_{k \neq k_0} p_k Q_k + Q_{k_0}\right) \eta(S_1 + 1) p_{k_0} \\ \text{subject to} \quad & p_k = \{0, 1\}, \quad \forall k \neq k_0 \\ & p_k = S - S_1, \quad k = k_0 \end{aligned}$$

which is a combinatorial problem and the solution is simply given by $p_{\pi(k)} = 1,\ 1 \leq k \leq \lfloor S \rfloor$, $p_{\pi(k_0)} = S - \lfloor S \rfloor,\ k_0 = \lfloor S \rfloor + 1$, and $p_{\pi(k)} = 0$, otherwise.

Combining the discussions on Case 1 and Case 2, we obtain the results for $\{p_k\}$. In addition, using such results, we also obtain $\overline{R}_k = \mathbb{E}\left[\chi_k \eta(s)\right]$ as in Theorem 2.

APPENDIX D
PROOF OF LEMMA 3

Taking derivative of the objective function of $U(S)$, we obtain

$$\frac{d}{dS} U(S) \quad = \quad -\sum_{k=1}^{\lfloor S \rfloor} Q_{\pi(k)} \eta_{\pi(k)}(\lfloor S \rfloor) + \sum_{k=1}^{\lfloor S \rfloor + 1} Q_{\pi(k)} \eta_{\pi(k)}(\lfloor S \rfloor + 1) - V.$$

It is observed that, given any integer S_0, the gradient $\frac{d}{dS} U(S)$ remains constant for any $S \in (S_0, S_0 + 1)$. If $\frac{d}{dS} U(S) = 0$, we can consider S_0 or $S_0 + 1$ to be the local maximum. If $\frac{d}{dS} U(S) \neq 0$, using the optimality condition [21], $S \in (S_0, S_0 + 1)$ cannot be the maximum. It concludes that, the maximum should be an integer.

APPENDIX E
PROOF OF THEOREM 3

In the following, we use notations s and S interchangeably. Define a continuous function $\eta_c(s) = \eta(s)$. Unlike $\eta(s)$, we allow $\eta_c(s)$ to be defined on all positive real numbers. Let $\mathcal{I}$ denote the space of positively increasing concave functions, i.e.,

$$\mathcal{I} \triangleq \left\{ \phi \in \mathcal{C}^2(0, +\infty) : \phi > 0, \phi^{'} \geq 0, \phi^{''} \leq 0 \right\}.$$

Given $g \in \mathcal{I}$, define $G(s) = g(s)\eta_c(s) - Vs$. We have the following result.

Lemma 7: A sufficient condition for the problem (20) having a unique local maximum in $s = \{1, \ldots, K\}$ is that there is a unique local supremum of $G(s)$ in $s \in (0, K)$ for any $g \in \mathcal{I}$. ■

Proof: Note that, since $\{Q_{\pi(k)}\}$ are on decreasing order, $g_Q(S) = \sum_{k=1}^{S} Q_{\pi(k)}$ is increasing, but the increment $g_Q(S+1) - g_Q(S)$ is decreasing. Therefore, given any $\{Q_{\pi(k)}\}$, there exists a function $g \in \mathcal{I}$, such that $g(s) = g_Q(s)$, for $s = \{1, \ldots, K\}$. As such, the function $G(s)$ for $s \in (0, K)$ is an interpolation of the objective function $U(s)$, where $G(s) = U(s)$ for $s = \{1, \ldots, K\}$.

Therefore, it is straight-forward that if $G(s)$ has a unique local supremum for $s \in (0, K)$, so does $U(S)$ for $s = \{1, \ldots, K\}$. ■

To show $G(s)$ has a unique local supremum, it suffices to show $G(s)$ is concave, which is equivalent to show that

$$G^{''}(s) = g^{''}(s)\eta_c(s) + 2g^{'}(s)\eta_c^{'}(s) + g(s)\eta_c^{''}(s) \leq 0.$$

From the property of $g \in \mathcal{I}$, we have $g^{'}(s)s \leq g(s)$. Thus

$$G^{''}(s) \leq g^{''}(s)\eta_c(s) + \frac{g(s)}{s}\left[2\eta_c^{'}(s) + s\eta_c^{''}(s)\right]. \tag{35}$$

The first term is negative by the definition of $g \in \mathcal{I}$. In the second term, $\frac{g(s)}{s}$ is positive. Now, let $\Gamma(s) = 2\eta_c^{'}(s) + s\eta_c^{''}(s)$. Note that, from (15), $\eta_c(s)$ is twice differentiable on $s \in (0, +\infty)$, and we have the following two equations

$$\eta_c^{'}(s) = M\int_0^{\infty} \log(1+x)N^2 f(x)\log[F(x)]F(x)^{NS-1}dx,$$

and

$$\eta_c^{''}(s) = M\int_0^{\infty} \log(1+x)N^3 f(x)\log\left[F(x)\right]^2 F(x)^{NS-1}dx.$$

Note that, for $N = 1$, $\Gamma(s; N = 1) \to 0$ as $s \to \infty$. It is also easy to verify that, $\Gamma(s; N = 1) \leq 0$ for all $s > 0$. For $N > 1$, let $t = Ns$. From the above two equations, we have $\Gamma(s; N) = N^2\Gamma(t; N = 1) \leq 0$. With $\Gamma(s) \leq 0$, we have $G^{''}(s) \leq 0$ in (35). Hence $G(s)$ is concave and has a global supremum in $(0, K)$. Using Lemma 7, we have proven the result.

APPENDIX F
PROOF OF LEMMA 5

Consider an upper bound ordered queue length profile as follows,

$$\hat{Q}_{\pi(1)} = Q_{\max}, \quad \hat{Q}_{\pi(j)} = Q_{\max}(1 - \delta\frac{j-1}{K})$$

where $\delta \geq 0$ is chosen such that $Q_{\pi(j)} \leq \hat{Q}_{\pi(j)}$ for all $j = \{1, \ldots, K\}$.

Note that in the outer subproblem (20), the term is increasing $\sum_{k=1}^{S} Q_{\pi(k)}$ with S while the term $\eta_{\pi(k)}(S)$ is decreasing. Hence using the upper bound queue length $\hat{Q}_{\pi(k)}$ yields an upper bound solution $\hat{S}^*$ to the outer subproblem (20).

We first solve the outer subproblem (20) using the upper bound queue lengths $\hat{Q}_{\pi(k)}$ approximated above and the approximated average data rate $\overline{R}_k = p_k\eta_k(S)$ from Lemma 4. The problem now becomes

$$\max_{S} \quad g(S) = \frac{Q_{\max}}{2K}(2K+\delta-\delta S)M\log\left(P\log NS\right) - VS. \tag{36}$$

It can be shown that $g(S)$ is concave. Taking derivative of $g(S)$, we obtain,

$$\begin{aligned} g^{'}(S) &= -\delta\frac{Q_{max}}{2K}M\log\left(P\log NS\right) \\ &\quad + \frac{MQ_{\max}}{2K}\left(2K+\delta-\delta S\right)\frac{1}{S\log NS} - V. \end{aligned}$$

Setting $g^{'}(S)=0$, we have

$$S\log NS = \left[\frac{V}{MQ_{\max}} + \frac{\delta}{2K}\left(\log\left(P\log NS\right) + \frac{1}{\log NS} - \frac{1}{S\log NS}\right)\right]^{-1}.$$

Therefore, we have

$$NS\log NS \le \left(\frac{V}{MQ_{\max}}\right)^{-1} N = \frac{MNQ_{\max}}{V} \triangleq c_1$$

for $S \ge 3$ and all $\delta \ge 0$. Thus we can just take $\hat{S}^* \le \frac{1}{N}e^{W(c_1)}$. Note that, under $\delta \to 0$, we have

$$\frac{\delta}{2K}\left(\log\left(P\log NK\right) + \frac{1}{\log NK} - \frac{1}{S\log NS}\right) \to 0,$$

which means the upper bound is achieved when $Q_{\pi(k)} \approx Q_{\max}$.

As a result, we have $S^* \le \hat{S}^* \le \frac{1}{N}e^{W(c_1)}$.

APPENDIX G
LARGE DEVIATION WITH SAMPLE PATH ANALYSIS

In this section, we give a brief introduction to the sample path large deviation analysis and derive the proof of Theorem 4.

A. The Large Deviation Principle

Consider the scaled sample path $q^B_{\max}(t) = \frac{1}{B}Q_{\max}\left(\lfloor Bt\rfloor\right)$, where the jumps can be given by[2]

$$q^B_{\max}(t) - q^B_{\max}(s) = \frac{1}{B}\sum_{\tau=\lfloor Bs\rfloor}^{\lfloor Bt\rfloor} A_{m(\tau)}(\tau) - \frac{1}{B}\sum_{\lfloor Bs\rfloor}^{\lfloor Bt\rfloor} D_{m(\tau)}(\tau)$$

for $0 \le s < t \le T_s$, where $m(\tau) = \arg\max Q_k(\tau)$ denotes the index of the maximum queue at time τ. Note that (at least when $|t-s|$ is small), the jump $q^B_{\max}(t) - q^B_{\max}(s)$ is a sum of independent variables

[2]Here, for easy discussion, we assume the identity $q^B_{\max}(\tau+\frac{1}{B}) - q^B_{\max}(\tau) = \frac{1}{B}A_{m(\tau)} - \frac{1}{B}D_{m(\tau)}$ holds on the boundary, where the maximum queue index changes, i.e., $m(\tau) \ne m(\tau+\frac{1}{B})$. Note that, with the fluid approximation, such boundary effect (which violates the above equality) vanishes in the scaled sample path $q^B_{\max}$ when B becomes large (and hence the jumps becomes smaller).

$v(\tau) = A_{m(\tau)} - D_{m(\tau)}$. Consider $q^B_{\max}(t)$ follows a continuous sample path $w(t)$, as $q^B_{\max}(t) \approx w(t)$. We write the random jump as $v(w(t))$, since it depends on the state $w(t)$.

Define the function $I_0 =$

$$\inf\left\{\int_0^{T_s} l(w(\tau), w^{'}(\tau))d\tau : w(0) = 0, w(T_s) = 1, T_s > 0\right\} \tag{37}$$

where

$$l(x = w(\tau), y = w^{'}(\tau)) = \sup_{\theta}\{\theta y - g(x,\theta)\} \tag{38}$$

is the local rate function, and

$$g(x,\theta) = \log \mathbb{E} e^{\theta v(x)}$$

is the *logarithm moment generating function* (LMF) of the random variable $v(x)$.

We consider the escape time $\tau_B = \inf\left\{t > 0 : q^B_{\max}(t) > 1\right\}$, when the process goes overflow following. Therefore, the following theorem characterizes the large deviation principle for τ_B.

Theorem 5: (Large deviation principle for τ_B) Suppose $\log v(x)$ is Lipschitz continuous on $[0, 1]$. We have the following,

(i) For each $\epsilon > 0$, $\lim_{n\to\infty} \Pr\left(\frac{\log \tau_B}{\tau_B} \in (I_0 - \epsilon, I_0 + \epsilon)\right) = 1$,

(ii) $\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}[\tau_B] = I_0$,

where $I_0 = \inf\{I^w : w(0) = 0, w(T_s) = 1, T_s > 0\}$. ■

Proof: Please refer to [15, Theorem 6.17] for the proof. ■

Note that the escape time τ_B that the process $q^B_{\max}(t)$ takes to enter the set $\{q^B_{\max}(t) > 1\}$ implies the steady state probability for $q^B_{\max}(t)$ to be in the set $\{q^B_{\max}(t) > 1\}$, i.e., $\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}[\tau_B] = \lim_{B\to\infty} -\frac{1}{B} \log \Pr\left(q^B_{\max}(\infty) > 1\right)$. The above theorem guarantees that $I^* = I_0$, and we can find the large deviation decay rate by solving 37. In the next section, we shall find the LMF $g(x,\theta)$ to solve I_0.

B. The Rate Function

We first characterize $g(x,\theta)$. Note that the random variables $v(t)$ is a composition of the arrival $A_{m(t)}$ and departure $D_{m(t)}$, which are independent. Therefore, we have $g(x,\theta) = \log \mathbb{E}\left[e^{\theta(A(x) - D(x))}\right] = \log \mathbb{E}\left[e^{\theta A}\right] + \log \mathbb{E}\left[e^{-\theta D(x)}\right]$. We find the LMF of the random variable $D(x)$ using the result of the following lemma.

Lemma 8: Suppose there are S users feedback to the BS. Then the data rate for user $k \in \mathcal{F}$ satisfies $R_k = lr$, where $r \to \log(P \log NS)$ almost surely (a.s.), and $l \to \xi \sim \text{Poiss}(\rho)$ in distribution, as $\rho = \frac{M}{S}$ and $S \to \infty$. ■

Proof: Let $r(S)$ be the data rate for the i-th beam under S users feedback to the BS. The results in [2] give that $\mathbb{E}\left[r(S)\right]/\log\left(P\log NS\right)\to 1$, as $S\to\infty$. Also, due to the channel hardening effect [22], the variance $\mathrm{Var}\left[r(S)\right]\to 0$, as $S\to\infty$. This shows that $r(S)\to\log\left(P\log NS\right)$, a.s.

According to the scheduling policy, each user $k\in\mathcal{F}$ may be allocated with $l=0,\ldots,\min\left\{M,N\right\}$ beams, and the data rate can be written as $R_k=lr(S)$. Since $\mathrm{SINR}_{k,n}^i$ are i.i.d. over k and n, the probability that a user being assigned l beams approximately follows a binomial distribution $B(M,p)$, where $p=\frac{1}{S}$.

It is well-known that $B(M,p)\to\mathrm{Poiss}\left(\rho\right)$, as $M\to\infty$ and $Mp\to\rho$. Therefore, the result is proven. ■

Using the above result, the package departure $D(x)=R_k/L\approx\xi\mu_0$ for $k\in\mathcal{F}$, where $\mu_0=r(S(x))/L$ and $\xi\sim\mathrm{Poiss}\left(\frac{M}{S(x)}\right)$. We have

$$g_D(x,\theta)=\log\mathbb{E}\left[e^{-\theta\xi\mu_0}\right]=\mu_0\log\mathbb{E}\left[e^{-\theta\xi}\right]=\mu_0\frac{M}{S}\left(e^{-\theta}-1\right).$$

Therefore, the LMF of $v(x)$ is given by $g(x,\theta)=g_A(\theta)+\mu(x)\left(e^{-\theta}-1\right)$, where $\mu(x)\triangleq\mu_0(S(x))\frac{M}{S(x)}$.

We make use of the following lemma to prove the final result.

Lemma 9: Assume that $l(x,y)$ in (38) is differentiable in y at all x, which is nondegenerate in $[0,1]$. For each x, the equation $g(x,\theta)$ has at most two solutions. Then with the appropriate choice of $\theta^*(w)$, we have

$$I^*=\int_0^1\theta^*(x)dx.$$

■

Proof: Please refer to [19, Lemma C.9] for the proof. ■

Using the result of the above Lemma, we prove the theorem.

APPENDIX H
PROOF OF COROLLARY 1

Under the CSI-only algorithm, the average data rate for each user k is given by $\overline{R}_{k,\mathrm{baseline}}\approx\frac{M\log(P\log NK)}{K}$, where the numerator is from Lemma 4 given all the K users feedback to the BS, and the denominator is due to the fact that on average all the users have equal probability to get scheduled. Therefore, the packet departure rate of the maximum queue $Q_{\max}(t)$ is given by $\mu(x)=\frac{\overline{R}_{k,\mathrm{baseline}}}{L}\approx\frac{M\log(P\log NK)}{KL}=\mu_b$.

Under the Poisson arrivals, we have $g_A(x,\theta)=\log\Lambda_A(\theta)=\lambda(e^\theta-1)$. Thus solving $g(x,\theta)=\lambda(e^\theta-1)+\mu(x)\left(e^\theta-1\right)=0$, we obtain $e^\theta=1$, or $\mu(x)/\lambda$. It is easy to verify that $e^\theta=1$ only leads to $I_0=0$, which is not in our interest. On the other hand, using $e^\theta=\mu(x)/\lambda$, we obtain the

large deviation decay rate function from Theorem 4 as

$$\begin{aligned} I^* &= \int_0^1 \log \frac{\mu(x)}{\lambda} dx \\ &= \int_0^1 \log \frac{M \log\left(P \log NK\right)/KL}{\lambda_T/K} \\ &= \log \frac{M \log\left(P \log NK\right)}{\lambda_T L}. \end{aligned}$$

APPENDIX I
PROOF OF COROLLARY 2

Using the result in Lemma 4, the average packet departure rate is given by

$$\mu_p(Q_{\max}) = \frac{\overline{R}_k}{L} \approx \frac{M \log\left(P \log N \hat{S}^*(Q_{\max})\right)}{L \hat{S}^*(Q_{\max})}.$$

Note that such approximation may be loose when $Q_{\max}$ is small. Instead, we use the following augmented approximation,

$$\widetilde{\mu}_p(Q_{\max}) = \max\left\{\mu_p(Q_{\max}), \frac{MR_0}{LK}\right\}$$

where $R_0 = \int_0^\infty \log(1+Px) dF(x)$. Note that R_0 is the average link capacity of a SISO channel and MR_0 is the average capacity under the random beamforming. Hence $\frac{MR_0}{LK}$ is the average package departure rate for the maximum queue process $Q_{\max}(t)$ under a round-robin scheduling policy, which performs as a lower bound for the proposed user scheduling algorithm.

Note that $\mu_p(x)$ is monotonically increasing. Define ϵ_K as the solution to $\mu_p(x) = \frac{MR_0}{LK}$, and $\epsilon = \inf\{\epsilon_K : K \geq K_0\}$ for some $K_0 < \infty$. Using Theorem 4, we have

$$\begin{aligned} I^* &\geq \int_0^1 \log \frac{\widetilde{\mu}_p(x)}{\lambda} dx \\ &= \int_0^1 \log\left(\frac{1}{\lambda_T/K} \max\left\{\frac{M \log\left(P \log N \hat{S}^*(x)\right)}{L \hat{S}^*(x)}, \right.\right. \\ &\qquad \left.\left. \frac{MR_0}{LK}\right\}\right) dx \\ &= \log \frac{M}{\lambda_T L} + \int_0^\epsilon \log R_0 dx \\ &\qquad + \int_\epsilon^1 \log \frac{\log\left(P \log N \hat{S}^*(x)\right) K}{\hat{S}^*(x)} dx \\ &= \log \frac{M}{\lambda_T L} + \epsilon \log R_0 + (1-\epsilon) \log K \\ &\qquad + \int_\epsilon^1 \log \frac{\log\left(P \log N \frac{1}{N} \exp\left(W(\frac{MNx}{V})\right)\right)}{\frac{1}{N} \exp\left(W(\frac{MNx}{V})\right)} dx \\ &= \log \frac{M}{\lambda_T L} + \epsilon \log R_0 + (1-\epsilon) \log K + C \end{aligned}$$

where $C = \int_{\epsilon}^{1} \left\{ \log \left[N \log \left(PW \left(\frac{MNx}{V} \right) \right) \right] - W \left(\frac{MNx}{V} \right) \right\} dx$.

The first inequality is because $\widetilde{\mu}_p(Q_{\max})$ is a lower bound approximation of the packet departure rate $\mu_p(Q_{\max}) \approx \frac{\overline{R}_k}{L}$. Thus the corollary is proven.

REFERENCES

[1] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[2] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[3] P. Xia and G. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1853 – 1863, may 2006.

[4] J. Zheng and B. Rao, "Capacity analysis of MIMO systems using limited feedback transmit precoding schemes," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2886 –2901, july 2008.

[5] A. Bayesteh and A. Khandani, "On the user selection for MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1086–1107, 2008.

[6] W. Zhang and K. Letaief, "MIMO broadcast scheduling with limited feedback," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1457–1467, 2007.

[7] S. Sanayei and A. Nosratinia, "Opportunistic downlink transmission with limited feedback," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4363–4372, 2007.

[8] J. Diaz, O. Simeone, and Y. Bar-Ness, "Asymptotic analysis of reduced-feedback strategies for MIMO Gaussian broadcast channels," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1308–1316, 2008.

[9] Y. Cui, Q. Huang, and V. Lau, "Queue-aware dynamic clustering and power allocation for network MIMO systems via distributed stochastic learning," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1229 –1238, march 2011.

[10] F. She, W. Chen, H. Luo, and D. Yang, "Joint queue control and user scheduling in MIMO broadcast channel under zero-forcing multiplexing," *International Journal of Communication Systems*, vol. 22, no. 12, pp. 1593–1607, 2009.

[11] D. Djonin and V. Krishnamurthy, "MIMO transmission control in fading channels - a constrained Markov decision process formulation with monotone randomized policies," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 5069 –5083, oct. 2007.

[12] F. Fu and M. van der Schaar, "Decomposition principles and online learning in cross-layer optimization for delay-sensitive applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1401 –1415, Mar 2010.

[13] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.

[14] K. Huang and V. Lau, "Stability and delay of zero-forcing SDMA with limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6499 – 6514, Oct 2012.

[15] A. Shwartz, A. Weiss, and R. Vanderbei, *Large deviations for performance analysis*. Citeseer, 1995, vol. 107.

[16] J. Chung, C. Hwang, K. Kim, and Y. Kim, "A random beamforming technique in MIMO systems exploiting multiuser diversity," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 848–855, 2003.

[17] J. D. C. Little, "A proof for the queuing formula: L= λ w," *Operations Research*, vol. 9, no. 3, pp. 383–387, May 1961.

[18] M. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 2915–2934, 2006.

[19] A. Weiss, *Large deviations for performance analysis: queues, communications, and computing*. Chapman & Hall/CRC, 1995.

[20] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[22] B. Hochwald, T. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1893–1909, 2004.